Epitaph or Breaking News? Analyzing and Predicting the Stability of Knowledge Base Properties

Ioannis Dikeoulias
Saarland University
Saarbrücken, Germany
s9iodike@stud.uni-saarland.de

Jannik Strötgen
Bosch Center for Artificial Intelligence
Renningen, Germany
jannik.stroetgen@de.bosch.com

Simon Razniewski
Max Planck Institute for Informatics
Saarbrücken, Germany
srazniew@mpi-inf.mpg.de

ABSTRACT

Knowledge bases (KBs) contain huge amounts of facts about entities, their properties, and relations between them. They are thus the key asset in any intelligent system for tasks such as structured search and question answering. However, due to dynamics in the real world, properties and relations change over time, and stored knowledge may become outdated. While KB information evolves steadily, there is no information whether or not a KB property might be subject to change with high probability or whether it is likely to be stable. Systems exploiting KB information, however, could benefit a lot if they had access to this kind of information. In this paper, we analyze and predict the stability of KB entries, which allows to accompany entries with stability scores. Our predictive model exploits entity-based features and learns through historic data. A particular challenge to determine stability scores is that KB entries are not only added or modified due to real-world changes but also to reduce the incompleteness of KBs in general. Nevertheless, our evaluation of sample properties demonstrates the effectiveness of our method for predicting the one-year stability of KB properties.

KEYWORDS

knowledge bases; temporal validity; stability prediction

ACM Reference Format:
Ioannis Dikeoulias, Jannik Strötgen, and Simon Razniewski. 2019. Epitaph or Breaking News? Analyzing and Predicting the Stability of Knowledge Base Properties. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3314998

1 INTRODUCTION

Motivation. Knowledge bases like Wikidata (WD) are cornerstones of the Semantic Web, and essential for tasks such as entity linking [14], structured search (in particular entity search [3]), and question answering [5]. They usually continuously evolve, by regularly recrawling sources (e.g., DBpedia or YAGO), or by manual editing (e.g., Wikidata), and hence, what is not in them today may be in them tomorrow. Consequently, the question about the extent to which data can be trusted over time evolves. While it has been shown that explicit and implicit temporal information contained in knowledge bases (in combination with temporal tagging of involved texts [15]), can be exploited for entity linking [1], question answering [9], and information retrieval [4], the question about the stability of KB entries has not been raised.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3314998

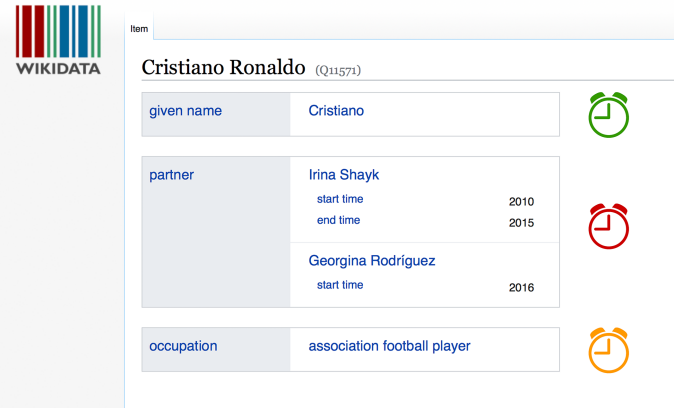


Figure 1: Envisioned visualization of stability information in Wikidata.

Consider for instance Cristiano Ronaldo, 34-year old soccer player currently dating model Georgina Rodriguez. What is the likelihood that the information about his name, occupation, and partner will still be valid in two years?

Contributions. In this paper, we focus on the stability aspect of *subject-property pairs* in Wikidata, like *(Ronaldo, occupation)*. We describe temporal information in Wikidata, characterize and measure the possible changes to subject-property-pairs, and build a predictive model for stability prediction of KB properties. Our evaluation on sample properties of humans shows promising results of our model and improvements over baselines.

The results of our approach are useful for better utilizing KBs in information retrieval settings such as structured search and question answering, but also for NLP tasks (e.g., entity resolution), for understanding KB quality, and for informing and guiding KB editors and readers, as exemplarily shown in Figure 1.

2 BACKGROUND

Related Work. Understanding temporal changes is crucial to estimate KB completeness [13]. A few works have analyzed how to anticipate or manage KB changes, for instance by monitoring relevant news sources [10], by analyzing verb changes in Wikipedia

articles [18], or by devising update strategies based on a cost-benefit analysis [12]. In [11], the stability of textual descriptions of Wikidata properties was analyzed. An approach based on association rules to predict parts of knowledge bases that are complete at a fixed timepoint is presented in [7]. Approaches to exploit Wikipedia edits for gaining insights about event updates and temporal information retrieval are described in [8] and [17], respectively.

Temporal information in Wikidata. Some properties in WD provide explicit temporal information, for instance *date of birth* and *date of death*. To other properties, temporal information can be added via so-called qualifiers, most importantly *point in time*, *start time* and *end time*, allowing to express, for instance, that Rodriguez is Ronaldo's partner since 2016 (see Fig. 1). Yet, in general, these are scarce, existing only for around 2-3% of statements.[1] A third source of temporal information is the WD revision history, which record the timepoint of the creation and modification of each statement.

3 ANALYZING KB MODIFICATIONS

For our analysis, we compared two Wikidata dumps from January 2017 and January 2018. We then identified the cases where subject-predicate pairs observed a change in frequency. In this work, the focus lies on the analysis of properties about humans, finding, for instance, for human subjects and the property *child*, a total of 21,328 changes. From all changes for humans, 89% were cases where objects for a property were newly created, 7% were cases where additional objects were added, 3% were deletions of all objects, and 1% were reductions.

Types of Property Changes. To get a better understanding of the types of changes of the properties about humans, we analyzed the 20 most common properties regarding whether they were newly added or subject to an increase.

As shown in Figure 2, *occupation* (+2.2%), *award received* (+1.2%), and *educated at* (+0.7%) were those observing the most increases (i.e., additional information was added). In contrast, *gender* (+11.8%), *occupation* (+9.8%), and *date of birth* (+8.5%) were most frequently newly created. We also identified a positive correlation of changes with the number of existing properties per entity, and a bell-shaped relation with the number of existing values per predicate (peak=5 values).

Reasons for Property Changes. Intuitively, the age of a person plays a big role regarding the likelihood that a property of a human changes. However, the reason for changes of KB properties can be two-fold: 1) real-world changes ("breaking news"), and 2) completion of old facts ("epitaph"). In the following, we discuss one Wikidata property with explicit time information (*child*) and one property with time information via qualifiers (*academic degree*).

Figure 3 shows the frequency of real-world changes that are reflected in the KB within the one year of our analysis, versus the changes in the KB that correspond to facts that happened earlier in reality grouped based on the age of a human. As one can see, the two sets of curves are very different. Changes in the real world happen predominantly between the age of 20 and 50, whereas change rates in the KB steadily increase with the age of persons, peaking at an age of 90, and only gradually decreasing beyond that age.

[1] According to our analysis of Wikidata dumps from 2017 and 2018.

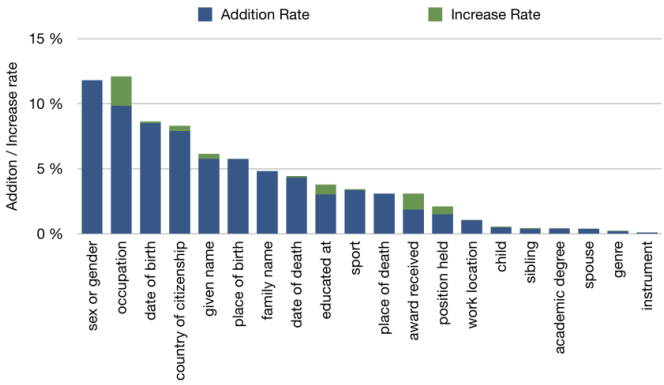


Figure 2: Addition and change rate of most frequent human properties.

This highlights that KBs predominantly lag behind with the real-world, and add data much later in life, e.g., only once persons become famous, or even upon death. This observation makes it particularly challenging to predict changes of knowledge base facts as they do not only rely on changes in the real world – which would often already be challenging to predict [2].

4 PREDICTING STABILITY OF KB PROPERTIES

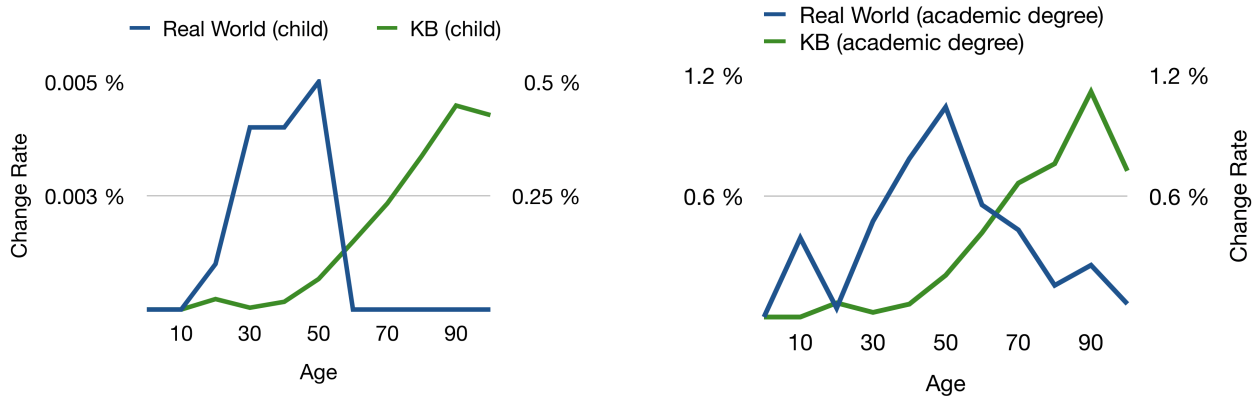
The second main goal of our work is to predict stability in knowledge bases. In particular, we looked at the following problem.

Problem 1 (Stability prediction). *Let e be an entity and p be a property. Given the current timepoint t and a future timepoint t', what is the probability that a fact in the set of all p-facts for e remains unchanged between t and t'?*

An instance of the problem is shown in the introduction: Assuming today is March 6, 2019, what is the likelihood that the *partner* information for Ronaldo will be unchanged on March 6, 2019? Note that for this problem, we can naturally only use information available as of March 6, 2018, and cannot continuously monitor textual sources, as done in [18] and [10].

Approach. What we can use is current information about the entity. In particular, we use the presence of properties and values. We proceed in three steps, the construction of feature vectors, model training, and model application.

(1) *Build entity-level feature vectors.* We construct binary feature vectors for entities by introducing 43 binary features that indicate for each of the 43 most common properties their presence or absence at time t. We also introduce binary features that, for each of the 43 properties, indicate the presence or absence of the 5 most frequent properties, if these occur at least 50 times. The rational is that for properties such as occupation, frequent values such as *politician* and *scientist* are possibly informative, whereas for properties such as child, entities rarely share objects at all. In the end, we thus introduce 149 additional features for object values.



(a) Child.

(b) Academic degree.

Figure 3: Real-time (left scale) vs. late (right scale) additions for two properties.

(2) *Train models based on historic data.* We next train predictive models using two historical Wikidata snapshots. Our target variables are again binary, one variable per property, indicating whether the respective property has changed between the two timestamps. In particular, we use logistic regression as model, because regression allows to learn probabilities even though given only binary values for the target variables.
(3) *Apply the models to individual entities.* Once having trained the models, we can apply them to predict future change probabilities. We feed a feature vector for a given entity to the regression model, and obtain as output a stability probability.

Experimental Setup. We verified the feasibility of this approach using the Wikidata dumps from January 2017 and 2018, on three properties, *child*, *academic degree* and *occupation*. In each case, we trained with 60% of the 3.7 million human entities in the 2017 dump, and used the remaining 40% for testing. A challenge for regression were correlations in our feature set, thus, we scaled down our feature set based on correlation analysis to just the 10 most distinctive features per target variable.

Baselines. We consider three baselines, namely to predict that exactly those persons that are *alive*, *dead*, or *active* (between 20 and 60 years old) are the ones that are going to observe change. The first baseline mirrors the expectation that only alive entities change and do so sufficiently often, while the second quantifies the reverse assumption, that entries of dead subjects are the focus of changes. The third represents a further refinement of the first, focusing on the typical age in which people are active.

Results. To systematically evaluate the performance of our approach, we translate the output probabilities back into binary *change/no change* predictions, and report precision, recall, and f-score in terms of the *change* class. Note that accuracy is not a useful measure, because, due to the high class imbalance, even classifiers that would entirely predict *no change* would achieve >98% accuracy. This becomes also evident via the baselines, which all obtain very low precision scores, due to predicting far too many changes. Furthermore, also their recall shows no clear preference between *alive* and *dead*, illustrating that changes are not confined to one of these sets.

Table 1: Change prediction results of the baselines (BL) and our prediction model (model).

	child			degree			occupation		
	P	R	F	P	R	F	P	R	F
BL-Alive	.00	.23	.00	.00	**.55**	.01	.13	**.71**	.22
BL-Dead	.00	**.77**	.01	.00	.45	.00	.04	.29	.07
BL-Active	.00	.03	.00	.00	.12	.00	.09	.20	.12
Model	**.05**	.12	**.07**	**.87**	.42	**.57**	**.51**	.43	**.47**

Our regression models, in contrast, achieved 5%, 87% and 51% precision and 12%, 42%, and 43% recall for the properties child, degree, and occupation, respectively. For degree and occupation, the high precision values are impressive when taking into account the rather low number of changes overall. In addition, the recall values are still on a promising level for these two properties.

Note that predicting future changes based on present states is truly difficult, because none of the individual features gives a truly strong indication for an upcoming change, and the class imbalance makes the prediction even harder. The child property seems to be particularly challenging probably because child information is often modified in Wikidata once a child becomes famous and thus a Wikidata entity.

Analysis. We can also inspect individual weights learned by the regression, finding for instance that the existence of children is a strong indicator for further changes in the *child* predicate (weight +9.0), while the existence of a father or siblings have a smaller negative influence (-0.5/-0.4). Interestingly, even the existence of *place of death* and *place of burial* have small positive weights w.r.t. change (+0.05/+0.13).

5 CONCLUSIONS AND ONGOING WORK

Our work is a first step towards understanding and predicting stability in knowledge bases. Interesting findings so far are that real-world and KB changes can occur very much disconnected, and that our model achieves promising results for the challenging task of stability prediction. Nevertheless, we are planning to extend this work along several dimensions:

(1) Features: Extending the idea from [18], we aim to include textual features, with the hope that these could indicate not only past, but also future change, and latent representations via knowledge graph embeddings [16], as well as features derived from temporal signatures for entites, as introduced in [1] for time-aware named entity disambiguation.
(2) Models: We plan to utilize other models, in particular, interpretable decision trees, and performant neural models.
(3) Change interrelation: We are currently working on analyzing the circumstances when these changes are connected or disconnected, in particular, whether certain changes co-occur, or follow particular order sequences, similar to the currency constraints in [6].

Furthermore, our goal is to deploy a prototype like shown in Figure 1 as plugin in Wikidata.

REFERENCES

[1] Prabal Agarwal, Jannik Strötgen, Luciano Del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. diaNED: Time-Aware Named Entity Disambiguation for Diachronic Corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL '18)*. 686–693. http://aclweb.org/anthology/P18-2109

[2] Ricardo Baeza-Yates. 2005. Searching the Future. In *Proceedings of the Mathematical/Formal Methods in Information Retrieval Workshop at SIGIR*.

[3] Krisztian Balog, Marc Bron, and Maarten De Rijke. 2011. Query Modeling for Entity Search Based on Terms, Categories, and Examples. *ACM Transactions on Information Systems (TOIS)* 29, 4, Article 22 (Dec. 2011), 22:1–22:31 pages. https://doi.org/10.1145/2037661.2037667

[4] Leon Derczynski, Jannik Strötgen, Ricardo Campos, and Omar Alonso. 2015. Time and information retrieval: Introduction to the special issue. *Information Processing & Management* 51, 6 (Nov 2015), 786 – 790. https://doi.org/10.1016/j.ipm.2015.05.002

[5] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open Question Answering over Curated and Extracted Knowledge Bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 1156–1165. https://doi.org/10.1145/2623330.2623677

[6] Wenfei Fan, Floris Geerts, Nan Tang, and Wenyuan Yu. 2013. Inferring Data Currency and Consistency for Conflict Resolution. In *2013 IEEE 29th International Conference on Data Engineering (ICDE '13)*. 470–481. https://doi.org/10.1109/ICDE.2013.6544848

[7] Luis Galárraga, Simon Razniewski, Antoine Amarilli, and Fabian M. Suchanek. 2017. Predicting Completeness in Knowledge Bases. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. ACM, 375–383. https://doi.org/10.1145/3018661.3018739

[8] Mihai Georgescu, Nattiya Kanhabua, Daniel Krause, Wolfgang Nejdl, and Stefan Siersdorfer. 2013. Extracting Event-Related Information from Article Updates in Wikipedia. In *35th European Conference on Information Retrieval (ECIR '13)*. 254–266. https://doi.org/10.1007/978-3-642-36973-5_22

[9] Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. TEQUILA: Temporal Question Answering over Knowledge Bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, 1807–1810. https://doi.org/10.1145/3269206.3269247

[10] Jiaqing Liang, Sheng Zhang, and Yanghua Xiao. 2017. How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI '17)*. 3749–3755. https://doi.org/10.24963/ijcai.2017/524

[11] Thomas Pellissier Tanon and Lucie-Aimée Kaffee. 2018. Property Label Stability in Wikidata: Evolution and Convergence of Schemas in Collaborative Knowledge Bases. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. 1801–1803. https://doi.org/10.1145/3184558.3191643

[12] Simon Razniewski. 2016. Optimizing Update Frequencies for Decaying Information. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*. 1191–1200. https://doi.org/10.1145/2983323.2983719

[13] Simon Razniewski and Gerhard Weikum. 2018. Knowledge Base Recall: Detecting and Resolving the Unknown Unknowns. *ACL SIGWEB Newsletter* Spring, Article 3 (June 2018), 3:1–3:9 pages. https://doi.org/10.1145/3210578.3210581

[14] Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (Feb 2015), 443–460. https://doi.org/10.1109/TKDE.2014.2327028

[15] Jannik Strötgen and Michael Gertz. 2016. *Domain-sensitive Temporal Tagging*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool.

[16] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI '14)*. 1112–1119. http://dl.acm.org/citation.cfm?id=2893873.2894046

[17] Stewart Whiting, Joemon Jose, and Omar Alonso. 2014. Wikipedia as a Time Machine. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. 857–862. https://doi.org/10.1145/2567948.2579048

[18] Derry Tanti Wijaya, Ndapandula Nakashole, and Tom Mitchell. 2015. "A Spousal Relation Begins with a Deletion of engage and Ends with an Addition of divorce": Learning State Changing Verbs from Wikipedia Revision History. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP '15)*. 518–523. https://doi.org/10.18653/v1/D15-1059